

BanSpEmo: a Bangla audio dataset for speech emotion recognition and its baseline evaluation

Babe Sultana^{1,2}, Md Gulzar Hussain^{3,4}, Mahmuda Rahman^{1,5}

¹Department of CSE, Faculty of Science and Engineering, Green University of Bangladesh, Dhaka, Bangladesh

²Department of CSE, Faculty of Science and Engineering, United International University, Dhaka, Bangladesh

³School of Software, Nanjing University of Information Science and Technology, Nanjing, Jiangsu, China

⁴School of Computer Science and Artificial Intelligence, Changzhou University, Changzhou, Jiangsu, China

⁵Department of ICT, Mohammadpur Preparatory School and College, Dhaka, Bangladesh

Article Info

Article history:

Received Jun 11, 2024

Revised Sep 28, 2024

Accepted Oct 7, 2024

Keywords:

Audio dataset

Bangla SER

Emotion classification

Machine learning

Speech emotion

ABSTRACT

Speech interfaces provide a natural and comfortable way for humans to communicate with machines. Recognizing emotions from acoustic signals is essential in audio and speech processing. Detection of emotion in speech is critical to the next generation of human-computer interaction (HCI) fields. However, a lack of large-scale datasets has hampered the progress of relevant research. In this study, we prepare BANSpEmo, a demanding Bangla speech emotion dataset consisting of 792 audio recordings totaling more than 1 hour and 23 minutes. The recordings feature 22 native speakers and each speaker uttered two sets of sentences representing six emotions: disgust, happiness, anger, sadness, surprise, and fear. The dataset consists of 12 Bangla sentences, each expressed in these six emotions. Furthermore, a series of investigations are carried out to assess the baseline performance of the support vector machine (SVM), logistic regression (LR), and multinomial Naive Bayes models on the BANSpEmo dataset presented in this study. The studies found that SVM performed best on this dataset, with an accuracy of 87.18%.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Md Gulzar Hussain

School of Software, Nanjing University of Information Science and Technology

Nanjing, Jiangsu, China

Email: gulzar.ace@gmail.com

1. INTRODUCTION

Speech is an essential and preferred way of communication for people. It's an important technique to convey emotions and plays a significant role in human-machine interactions. Speech emotion recognition (SER) research has received significant attention over the last few years due to its application in remote patient monitoring systems, robotics, the psychological assessment of people and many more [1]. While tremendous progress has been achieved in SER for widely used languages like English and Mandarin, there is still a significant deficit in resources and research committed to less commonly studied languages. Bangla, spoken by about 250 million inhabitants globally [2], is one of the underdeveloped languages in the field of SER. Although a significant amount of studies has been conducted in the area of textual data in the Bangla language in emotion and sentiment analysis—such as analyzing basic emotions [3] sentiment analysis in Bangla English Code-mixed text [4], [5] and emotion classification [6]. These efforts have greatly enhanced research understanding and insights into the Bangla language of textual data domain.

Detection of emotions from SER is a growing research topic due to its importance in the community, society, and commercial domains. In the realm of speech recognition (SR) and natural language processing (NLP), an extensive range of speech corpora has been created for multiple languages. Although there has been lots of research on SER for various languages, such as English [7], [8], Urdu [9], Chinese [10], Italian [11] and others [12], [13], there have been just a few efforts at developing SER dataset for Bangla. Table 1 shows some previously developed speech emotion recognition datasets and their limitations.

Table 1. Comparison of some previous speech emotion dataset in Bangla and other languages

Article	Year	Language	Contributions	Limitations
Paper [14]	2023	Bangla	A Speech Emotion dataset named KBES is developed of 900 recordings.	Number of recordings is limited and gender balance is not considered.
Paper [9]	2022	Urdu	The first Urdu Speech Emotion dataset is developed with 2,500 recordings.	In the dataset, the disgust emotion is hard to distinguish and with the disgust emotion, the accuracy is low.
Paper [15]	2022	Bangla	A Speech Emotion dataset named SUBESCO is developed with 7000 recordings.	Purely neutral sentences are uttered with different emotions which is difficult to express.
Paper [16]	2022	Bangla	A Speech Emotion dataset named BanglaSER is developed of 1467 recordings.	Number of sentences for uttering and the number of recordings is limited.
Paper [17]	2022	Bangla	A Speech Emotion dataset is developed of 452 recordings.	Number of recordings is limited, the annotation process is not explained and gender balance is not considered.
Paper [7]	2021	English	A new Speech Emotion dataset named LSSED is developed.	Number of data is limited, only 820. Also, the total length is only about 20 minutes.
Paper [18]	2021	Bangla	A Speech Emotion dataset named ABEG is developed.	Dataset is not publicly available, has only 3 classes, and the annotation process is not clear.
Paper [13]	2020	Spanish, Portuguese, German, French	A Speech Emotion dataset named CMU-MOSEAS is developed with 40,000 multi-modal samples.	Data samples are not balanced for the 4 languages. Gender balance is not considered also.
Paper [8]	2018	English	A new Speech Emotion dataset named RAVDESS is developed with 7356 recordings.	Have limited lexical variability due to the inclusion of only two statements.
Paper [12]	2016	English	A Speech Emotion dataset named EmoReact is developed with 1102 audio-visual clips.	Gender effects are not considered and the dataset is not gender balanced. Also number of clips is limited.
Paper [10]	2006	Mandarin	A new Speech Emotion dataset named MASC is developed with 25,636 recordings.	The dataset is not gender balanced as it contains the recordings of 23 female and 45 male Chinese speakers. Also, the traditional speaker verification and identification systems are limited for the dataset.

From Table 1 it can be observed that there have been few efforts to create datasets for SER in the Bangla language. Dhar and Guha [17] created a dataset designated as ABEG. They employed three emotional states: angry, happy, and neutral. There was no further description of their dataset, and the data is not accessible to the public. A team of academicians prepared a small discrete corpus of 160 sentences to test the speech-emotion identification system they proposed [18]. This dataset has 20 individuals who represented emotions such as happy, angry, sad, and neutral where perceptual evaluation was not available. Nevertheless, just three corpora for the task of emotion detection from speech in the Bangla language are publicly accessible now: SUBESCO [19], BanglaSER [16], and KBES [14]. Using these openly accessible Bangla language datasets, multiple publications have been published illustrating how to detect emotions in Bangla speech employing machine learning and deep learning methods. Different ensemble learning approaches are compared in multiple trials to show that they outperform typical machine learning techniques. The research findings show that ensemble learning

approaches can reach a great accuracy of 84.37%, which is achieved by utilizing the bootstrap aggregation and voting method. Sultana *et al.* [15] used the SUBESCO and RAVDESS [8] datasets to undertake cross-lingual investigations involving cross-dataset training, multi-dataset training, and transfer learning in English and Bangla. The suggested model demonstrated cutting-edge perceptual ability, with weighted accuracy (WA) of 86.9% for the SUBESCO and 82.7% for the RAVDESS. Hassan *et al.* [20] combines a one-dimensional convolutional neural network (CNN) and a long short-term memory (LSTM) framework to create a fully connected network for SER, comparing the performance of these two datasets. Islam *et al.* [21] combines transformed features from three separate methodologies—chroma short-time fourier transform, short-time fourier transform (STFT), and mel-frequency cepstral coefficient (MFCC)—and feeds them into a 3 dimensional CNN block to extract the features. The outputs are then processed by a bidirectional LSTM layer to classify Bangla speech emotions. In article from Sultana and Rahman [22], the researchers employed the grid search method with five-folded cross-validation for determining the best parameters for the support vector machines (SVM), random forest, and XGBoost algorithms. They discovered that choosing the most important features enabled machine learning models to achieve high levels of accuracy, equivalent to deep learning models. A recent study from Aziz *et al.* [23] presents a CNN-based approach for SER in Bengali, using MFCC features and data augmentation approaches. This method produced remarkable accuracies of 90% on the SUBESCO and 78% on the BanglaSER datasets.

Research on the detection of emotions in cross-linguistic speeches has demonstrated that systems trained on a single language dataset often perform poorly when evaluated on a separate language corpus, yielding lower accuracy rates than monolingual recognition rates. This performance gap emphasizes the importance of language-specific datasets for accurate emotion recognition. Over the past few years, there has been extensive investigation into SER in various languages [24], [25]. Despite having limited natural speech corpora [26], [27] or verified recorded emotional speech corpora [14], [16], [19] published for the Bangla language. Relevant linguistics resources for recognizing emotions are still inadequate. The SER system uses various approaches to classify and analyze audio files to find embedded emotions. The initial stage in its improvement is to generate a dataset for the targeted language which is one of the main goals of this research. The following is a summary of this work's main contributions:

- This research introduces BanSpEmo, a needed diversified Bangla dataset for emotion recognition from voice. It comprises 12 distinct sentences uttered by 22 native speakers to represent six desired emotions. The total duration is 1 hour and 23 minutes.
- This dataset enables more comprehensive simulations of real-world scenarios by increasing the lexical and sentence variability, allowing machine learning techniques and deep neural networks to grasp their pattern better.
- However, using the BanSpEmo dataset, this study compared the performance of three well-known algorithms: logistic regression (LR), SVM, and multinomial Naïve Bayes for Bangla voice emotion classification.
- This research also shows an investigation of these algorithms against a few well-known audio features to evaluate their efficiency in classifying emotions in Bangla speech. After analyzing the results, we discovered each algorithm's performance, showing useful information about their usability for SER tasks in the Bangla language.

A detailed description of the dataset is provided in section 2. The proposed research framework is explained in section 3. While the performance analysis of machine learning algorithms applied to this framework and also discusses the research findings in section 4. Overall discussion and insights into the future directions of this work is provided in section 5.

2. CORPUS DESCRIPTION

Speech is one of the modes of communication on various online platforms, such as Facebook and YouTube, where emotions are frequently conveyed. In this context, creating a speech dataset for the Bangla language is a significant contribution. The dataset we have prepared is available named BANSpEmo [28] constitutes the main portion of our research. As a low-resource language, Bangla has limited speech datasets. BANSpEmo marks the 4th audio dataset developed for SER in Bangla.

2.1. The experimental setup

The BANSpEemo dataset includes 792 voice recordings, capturing six fundamental emotional reactions across two sets of sentences, each with six sentences. The voices were recorded using a smartphone's recording application, a microphone, and a laptop. To create the dataset, we used our university's dedicated research lab, which was not entirely soundproof like a professional audio recording studio, but it was essentially noiseless. We took extra care to eliminate any background noise, including human or other ambient sounds. We made sure the recording environment was as controlled as possible by implementing stringent measures to reduce background noise. Maintaining these cautions, we were able to record audio that was both consistent and clear enough for our study. Each recording, lasting 5-6 seconds on average per emotion, had noise removed using Audacity software. Additionally, we used WavePad Sound Editor software for further editing. The summary of tools required for making the dataset:

- Microphone: BOYA BY-BM3011 compact shotgun microphone.
- Sound editor: WavePad sound editor.
- Audio noise remover: audacity software.

2.2. Corpus creation process

The speakers naturally conveyed the emotional states, ensuring that the recordings were not merely read aloud. The emotions represented are happiness, disgust, sadness, anger, surprise, and fear. This dataset focuses on data collected from individuals aged between 20 and 25. The corpus comprises voice recordings from 22 speakers, with an equal distribution of 11 males and 11 females. The duration of the tapes varies between 3 and 12 seconds, influenced by the length of the sentences and the time the speaker takes. While there are roughly equal numbers of male and female speakers overall, neither sentence set reflects this balance. With $6 \text{ sentences} \times 1 \text{ repetition} \times 6 \text{ emotions} \times 18 \text{ speakers}$ and $6 \text{ sentences} \times 1 \text{ repetition} \times 6 \text{ emotions} \times 4 \text{ speakers}$, the total number of recordings is 792 utterances. The complete audio dataset spans a total of 1 hr, 23 mins, and 12 secs. The following Table 2 presents a summary of the dataset.

Table 2. Dataset description table

Type of dataset	Performed, scripted
Type of File	Audio
Language	Bangla
Gender	Male and Female
Data format	Waveform Audio File Format (WAV)
Number of Groups	2
Number of Sentences per Group	6
States of Emotion	Happiness, Disgust, Sadness, Anger, Surprise, and Fear
Total Number of Statements	12
Total Number of Audio Tapes	792

2.3. Details of sentences

We selected 12 sentences to ensure diversity, as these sentences are typically used to express different emotions. We trained our speakers to deliver each selected sentence with six emotions to prepare our audio dataset. A wide range of emotional expressions, such as happiness, sadness, anger, fear, surprise, and disgust, were carefully considered when crafting each sentence. By doing this, we hope to build a solid dataset that will be useful for a range of speech emotion recognition and affective computing applications. The orators underwent comprehensive training to guarantee uniformity and precision in their emotive communication. The chosen Bangla text and their English meanings are shown in the Table 3.

In Table 4, we aim to present a comparison of existing freely accessible SER datasets in this domain alongside our work, BANSpEemo. Given the volume of audio recordings and the total duration, this collection ranks as the fourth-largest emotional speech database in the Bangla language. Despite its relatively small size compared to other datasets, and with participation levels being fairly typical, the key strength of this dataset is its broad range of sentence variations. This lexical and sentence diversity enhances the ability to capture diverse emotional expressions in different forms in Bangla speech. Essentially, we chose a wide variety of sentences to explore how different expressions of the same emotion can be introduced in Bangla speech. To systematize

future training standards, we divided our BANSpEemo dataset into the training and the test sets. Initially, we shuffled all samples and then allocated 20% to the test set, leaving 80% for the training set. It was made sure that the distribution of each emotion class in both the training and test sets was consistent or at the very least, similar.

Table 3. The selected Bangla text and their English meaning

SL.	Bangla sentence	English meaning
1.	কিছু তথ্য সঠিক ভাবে উপস্থাপন করা দরকার, বার বার একই ভুল করে চলেছে সংবাদ মাধ্যম গুলি।	Some information needs to be conveyed appropriately, and the media is making similar mistakes repeatedly!
2	আপনার ব্যবহার তো চমৎকার। মুখের ভাষা ও অনেক সুন্দর।	Your behavior is wonderful. Your words are also pleasant.
3	এর পরিপ্রেক্ষিতে শিক্ষকদের স্বার্থ সংশ্লিষ্ট শিক্ষক সমিতির মধ্য থেকে কোনো ধরনের ভূমিকা পরিলক্ষিত না হওয়ায় আমি ভীষন ভাবে উদ্ভিষ্ট।	In this regard, no role has been observed from the teacher's associations regarding the interest of teachers made me densely concerned.
4	আমার একটা ব্যাপার মাথায় ধরে না, "ইলিশ বাঁচাও" স্লোগান মুখরিত মিডিয়া কেন এবং কি কারণে "ইলিশের বাসস্থান (নদী) বাঁচাও" স্লোগান নিয়ে মাতে না?	Why the slogan "Save the habitat (river) of Hilsa" rather than "Save the Hilsa" is being avoided by the media baffles me.
5	দেশ কি মধ্যম আয়ের দেশে রূপান্তর হচ্ছে নাকি মগের মূলুকের দেশে পরিণত হচ্ছে?	Is the country turning into a middle-income country or a country of chaos?
6	আমি একমাত্র সরকারি কোন কাজে আঙ্গুলের চাপ দিতে রাজি আছি, শিক্ষিত ব্যক্তি আঙ্গুলের চাপ দেয় না।	I agree to have my fingerprints used for government purposes, but reasonable people might not.
7	তগো মনে কতো প্রেম রে! জীবনে একটা করছি তাতেই জ্বলে পুড়ে শেষ।	You are bursting with love! I once tried to embrace it, but I got burned.
8	আজকের ম্যাচ ভারতকে হারাতে চাই টাইগার বাংলাদেশ সাবাস সাকিব আল হাসান।	To defeat India in today's match, we need the tiger of Bangladesh, Well done Shakib Al Hasan!
9	টাইটানিক জাহাজ ডুবে গেছে আর বাংলাদেশ ও ডুবে যাবে।	The Titanic has sunk, and Bangladesh will sink too.
10	প্রশ্ন যদি ভুল হয় তাহলে পরীক্ষা নেবার কি দরকার? সবাইকে গড়ে প্লাস দিয়ে দিবে।	If the questions are incorrect, what's the sense of taking the exam? Simply give everyone the A+ grade.
11	যদি খায় পানতা ইলিশ জুতা দিয়ে তার গালটা কর মালিশ।	He should be punished for making extravagant expenses during the price hike of hilsa.
12	যে জাতি পঁচা ভাত খেয়ে বছর শুরু করে, এরা উন্নতি লাভ করবে কি করে!	A nation that starts the year by eating spoiled rice, how will they ever progress!

Table 4. A comparison between publicly available Bangla Language SER corpora and the BANSpEemo

Description	SUBESCO	BanglaSER	KBES	BANSpEemo
Audio Clips	7000	1467	900	792
Emotions	7	5	9	6
Sentences	10	3	N/A	12
Participant	20	34	35	22
Trained Actors	Yes	No	Yes	No
Rate of Sampling Rate	48 kHz	44.1 kHz	48 kHz	44.1 kHz
Class Equilibrium	Yes	Yes	Yes	Yes
Gender Equilibrium	Yes	Yes	Yes	Yes

3. METHOD

In this Figure 1, we present our proposed system architecture. The collected raw data underwent a thorough cleaning and preprocessing stage, with mel-frequency cepstral coefficients, spectrogram (MFCCs), zero crossing rate (ZCR), root-mean-square energy (RMSE), and chroma being utilized as a feature extraction technique. We have applied several well-known machine learning algorithms support vector machine, logistic regression, and multinomial Naive Bayes to provide a comparative performance evaluation of existing tasks.

3.1. Data cleaning and preprocessing

To augment the dataset, each audio is divided into three segments. In the data preprocessing and cleaning phase, every audio undergoes trimming, we remove portions where no voice is detected. These segments typically correspond to pauses or moments when the speaker takes a breath. Additionally, we have standardized the frequency of each split audio to 44.1 kHz to ensure uniformity across all instances. Subsequently, features are extracted from each trimmed audio.

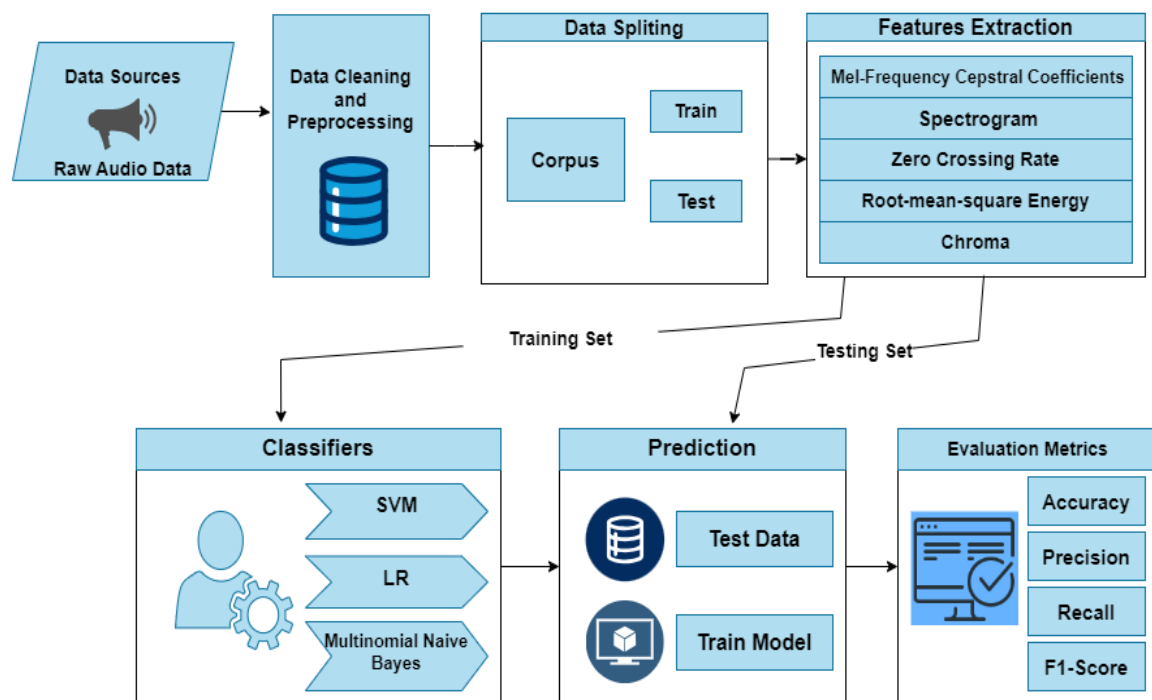


Figure 1. The suggested system's flow architecture

3.2. Feature extraction

3.2.1. Mel-frequency cepstral coefficients

To illustrate the short-term power spectrum of a voice signal, MFCCs are a set of coefficients broadly used in voice and audio processing. It is a condensed set of features, typically around 10 to 20. They serve as valuable features for machine learning models due to their ability to succinctly capture the key attributes of an audio signal while also reducing its dimensionality. In our feature extraction process, we calculated 20 MFCCs using the 'librosa.feature.mfcc()' Python module. Figure 2 illustrates the MFCC feature waveform for "Happy Emotion".

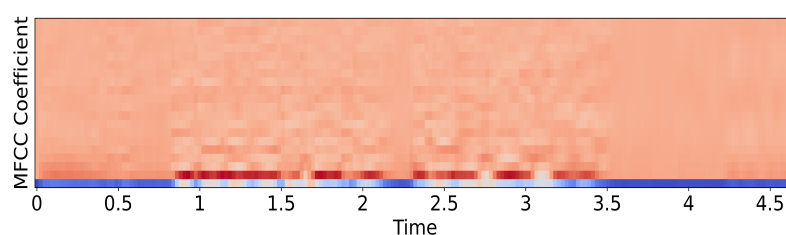


Figure 2. Sample MFCCs visualization for happy emotion

3.2.2. Spectrogram

A spectrogram behaves as a pictorial portrayal of how the frequency components of a signal evolve. It holds significant utility in signal processing, audio examination, and diverse scientific domains. Spectrograms provide a means to depict the alterations in signal frequencies across time, facilitating the examination and visualization of the evolving frequency characteristics of audio or other time-based signals. Figure 3 depicts a spectrogram visualization illustrating the signal's loudness over time across various frequencies in a specific waveform, for the "Happy Emotion".

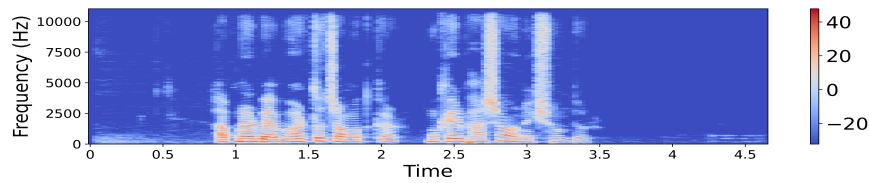


Figure 3. Sample spectrogram visualization for happy emotion

3.2.3. Zero crossing rate

In the domains of signal processing, audio analysis, and speech recognition, the ZCR is a frequently employed characteristic. It quantifies the speed at which a signal alters its polarity or intersects the zero amplitude line within a specified timeframe, essentially gauging how often a signal's waveform crosses the zero point. ZCR is formally defined as (1). Figure 4 illustrates the ZCR Visualization for the "Happy Emotion" which portrays the rate at which the signal transitions either from negative to zero to positive or from positive to zero to negative.

$$zcr = \frac{1}{T-1} \sum_{t=1}^{T-1} 1_{R<0}(s_t s_{t-1}) \quad (1)$$

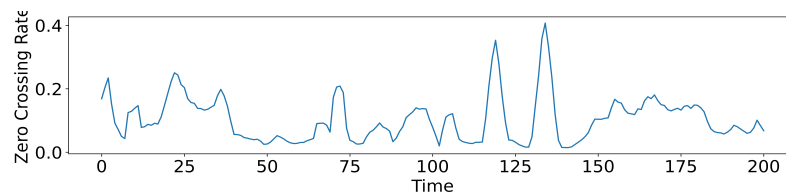


Figure 4. Sample ZCR visualization for happy emotion

3.2.4. Root-mean-square energy

In signal processing and diverse domains, RMSE is a mathematical metric employed to assess the energy level within a signal. It offers a means to characterize the amplitude or intensity of a signal within a defined time segment. The RMSE is defined as follows:

$$RMSE = \sqrt{\frac{1}{N} \sum_n |x(n)|^2} \quad (2)$$

here,

- RMSE is the root-mean-square energy.
- N is the number of samples in the time window.
- x(n) represents the signal samples.

In this context, with $N = 44,100$ and $x(n) = 204,800$. The RMSE value provides insight into the signal's energy and amplitude within the specified time frame, and Figure 5 depicts the visualization for "Happy Emotion".

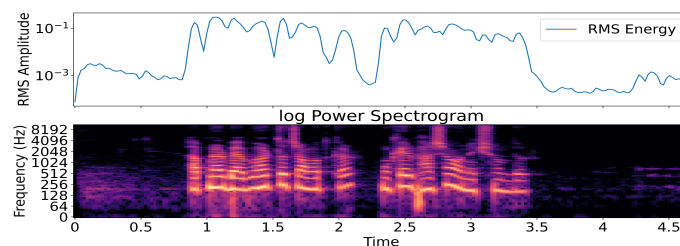


Figure 5. Sample RMSE visualization for happy emotion

3.2.5. Chroma

The chroma feature is a compact representation that conveys the tonal characteristics of a musical audio signal. Chroma features are derived from the chromagram representation of audio signals. These features encompass the chroma vector (depicting the intensity of each pitch class), chroma energy (the summation of squared chroma values), and chroma cross-correlation (which quantifies the similarity between chroma vectors). Figure 6 illustrates the visual representation of “Happy Emotion”.

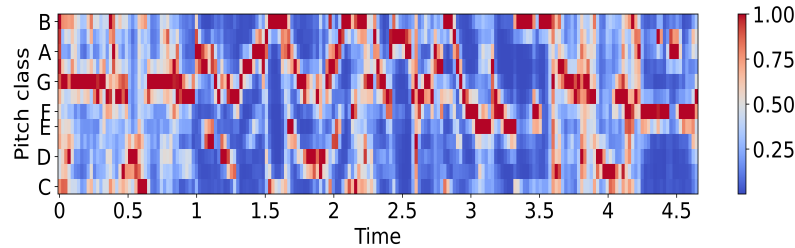


Figure 6. Sample chroma visualization for happy emotion

3.3. Classifier model

Following the data set cleaning, preprocessing, and feature extraction stages, we employed numerous machine learning algorithms. However, this research specifically focuses on three machine learning algorithms: SVM, LR, and multinomial Naive Bayes. These three algorithms were chosen due to their outstanding performance in terms of accuracy on this particular data set.

3.3.1. Support vector machine

The SVM stands out as a potent and adaptable machine learning algorithm extensively utilized in binary and multi-class classification tasks and regression. SVM functions by identifying the optimal hyperplane within the feature space, effectively maximizing the margin between distinct classes and facilitating accurate classification. This technique seeks to identify the optimal hyperplane for distinguishing between feature classes. One way to represent the equation for a linear SVM hyperplane is as:

$$f(x) = \arg \max_c (w_c \cdot x + b_c) \quad (3)$$

here,

- $f(x)$ is the decision function.
- c represents the different classes.
- w_c is the class c weight vector.
- x is the input feature vector.
- b_c is the class c bias term.

In this configuration, the class with the highest score from the decision function is chosen to determine the projected class for an input sample. This technique is frequently used in multiclass SVM settings, where different binary classifiers are trained using one of two strategies: one-vs-one (OvO) or one-vs-rest (OvR). We have employed the one-vs-rest (OvR) technique in our implementation. The final class assignment is subsequently determined by taking into account the outputs of each binary classifier’s decision function.

3.3.2. Logistic regression

The popular machine learning method known as LR was initially created for binary categorization. With the help of this technique, LR may be used effectively in scenarios with more than two classes, providing insightful information about the likelihood that each class would be an accurate prediction. But in this case, we’ve abandoned LR in favor of the OvR technique to handle multi-class classification problems. The OvR approach simplifies the LR adjustment for multi-class jobs, making it a versatile solution for various classification problems. The class c classifier models the probability that a z belongs to class c in the following way:

$$P(y = c|z) = \frac{1}{1 + e^{-(w_c^T z + b_c)}} \quad (4)$$

here,

- The probability is $P(y = c|z)$ which means the output y belongs to class c .
- For class c , w_c is the weight vector and b_c is the bias term.
- Vector of input features is z .

3.3.3. Multinomial Naive Bayes

Among the probabilistic classification techniques designed for discrete feature scenarios is the multinomial Naive Bayes, which is widely applied to text classification problems. Building on the foundations of Bayes' theorem, this method is designed assuming that the features, given the class label, exhibit conditional independence. Multinomial Naive Bayes extends its usefulness in a multi-class classification scenario by using the concepts of the Bayes theorem to determine the probability that an instance will be assigned to each class. The equation for predicting the class x probability is given the features $f_1, f_2, \dots, f_n$ can be shown as:

$$P(X = x|f_1, f_2, \dots, f_n) \propto P(X = x) \prod_{i=1}^n P(f_i|X = x) \quad (5)$$

Here,

- X is used as a class variable.
- $f_1, f_2, \dots, f_n$ is used as the feature variables.
- $P(X = x|f_1, f_2, \dots, f_n)$ is the class x posterior probability of given the features.
- $P(X = x)$ is the class x prior probability.
- $P(f_i|X = x)$ is the conditional probability of feature f_i given class x .

These probabilities are computed using the training dataset during the training step. During the prediction step, the algorithm then computes the succeeding probabilities for every class, identifying the class with the maximum probability as the forecasted class for the specified set of features. The algorithm's ease of use in handling multi-class classification problems can be ascribed to its effectiveness, ease of handling high-dimensional data, and simplicity.

4. PERFORMANCE EVALUATION

This section compares the outcomes of several machine learning algorithms and presents their performance analyses. It discusses the environmental setup, evaluation metrics analysis of different machine learning models, performance comparison with some previous works, confusion matrix analysis, and receiver operating characteristic (ROC) curve analysis.

4.1. Environmental setup

- Operating system: Windows 10 64 bit
- Processor: Intel(R) Core(TM) i5-4300M CPU @ 2.60 GHz
- RAM: 8 GB
- IDE : Google Colab
- Programming language: Python

4.2. Result analysis and discussion

Our goal in this section is to present a thorough analysis and discussion of the findings from the many assessment metrics we used in our study. We have also looked at ROC curves, which serve as an effective and lucid visual aid for illustrating the classifier's accuracy.

4.2.1. Evaluation metrics analysis

We employ a range of standard performance assessment metrics to evaluate and contrast the effectiveness of different classifiers. Our comparative analysis evaluates the relative performance of classifiers using accuracy, precision, recall, and F1 scores. Accuracy is utilized by comparing the predicted labels of each instance with the ground-truth labels, but its limitations are acknowledged as certain samples may introduce bias. Therefore, we also incorporate precision, recall, and F1 measures to provide a more comprehensive evaluation. We experimented with various machine learning algorithms and eventually selected three—SVM, LR, and multinomial Naive Bayes—due to their commendable performance in this context. Among these, SVM exhibited the highest accuracy at 87.18%, followed by LR at 84.45%, and multinomial Naive Bayes at 82.77%. In Tables 5-7, we presented the individual precision, recall, and F1 scores for all emotions considered in our

research. From these tables, it is evident that SVM attains the highest weighted average values for precision, recall, and F1 score, with 0.87, 0.87, and 0.86, respectively. In contrast, LR yields weighted average precision, recall, and F1 scores of 0.85, 0.84, and 0.84, respectively. For multinomial Naive Bayes, the corresponding values are 0.83, 0.83, and 0.82.

Table 5. Outcomes of SVM-based precision, recall, f1-score, and accuracy in six distinct categories (emotions)

Category	Precision	Recall	F1-Score	Accuracy (%)
Anger	0.92	0.93	0.92	87.18%
Disgust	0.82	0.88	0.85	
Fear	0.85	0.87	0.96	
Happy	0.88	0.93	0.90	
Sad	0.91	0.82	0.86	
Surprised	0.82	0.66	0.73	

Table 6. Outcomes of LR-based precision, recall, f1-score, and accuracy in six distinct categories (emotions)

Category	Precision	Recall	F1-score	Accuracy (%)
Anger	0.85	0.92	0.88	84.45%
Disgust	0.89	0.89	0.89	
Fear	0.85	0.87	0.86	
Happy	0.78	0.83	0.81	
Sad	0.82	0.81	0.81	
Surprised	0.89	0.60	0.71	

Table 7. Outcomes of multinomial Naive Bayes-based precision, recall, f1-score, and accuracy in six distinct categories (emotions)

Category	Precision	Recall	F1-Score	Accuracy(%)
Anger	0.82	0.91	0.86	82.77%
Disgust	0.84	0.87	0.85	
Fear	0.81	0.87	0.84	
Happy	0.86	0.85	0.86	
Sad	0.80	0.78	0.79	
Surprised	0.82	0.51	0.63	

4.2.2. Performance comparison with relevant Bangla datasets

In the Table 8, we aim to compare this work with previous studies that have focused on Bangla SER. Hassan *et al.* [20] primarily utilized two datasets: one in English, named RAVDESS, and another dataset SUBESCO for Bangla. Their proposed model, which integrated a 1D CNN with a fully convolutional network (FCN) layer, achieved 98.30% accuracy on the RAVDESS dataset and 98.97% on the SUBESCO dataset.

Table 8. Comparison with related works used Bangla speech emotion datasets

Article	Dataset	Features extraction techniques	Classifier	Accuracy
Paper [20]	SUBESCO	MFCC, ZCR, Mel-Spectrogram, Root Mean Square, etc	1D CNN + FCN layers	98.97%
Paper [15]	SUBESCO	CNN + TDF layer	DCTFB	86.9%
Paper [21]	SUBESCO	MFCCs + STFT + Chroma STFT	4CNN + TDF + Bi-LSTM	89.57%
Paper [29]	KBES	MFCC, STFT, Chroma STFT, CNN	TDF layer, Bi-LSTM, LSTM	71.67%
Paper [30]	SUBESCO, Ban-glaSER	CNN	KNN, AdaBoost, Bi-LSTM	90%
This Work	BanSpEmo	MFCC, Spectrogram, ZCR, RMSE	SVM, LR, MNB	87.18%

Using the dataset SUBESCO paper [15] utilized CNN and TDF features with DCTFB classifier and achieved an accuracy of 86.9%. Additionally, Islam *et al.* [21] used 3D CNN and bidirectional long short-term memory networks (Bi-LSTM) as models while working with the SUBESCO dataset. They achieved an accuracy

of 89.57% by using a variety of speech signal modifications, including MFCC, chroma STFT, and short-time fourier transform (STFT). Another article from Billah *et al.* [29] used the KBES dataset to classify speech emotions with MFCC, STFT, chroma STFT, and CNN features on Bi-LSTM and LSTM classifiers. However, it has a comparatively lower accuracy of 71.67%. Shruti *et al.* [30] primarily focuses on a comparative analysis using the SUBESCO and BanglaSER datasets. The authors applied KNN, AdaBoost, and Bi-LSTM algorithms, with KNN achieving an accuracy of 90.0%, while AdaBoost and Bi-LSTM reached only 45% accuracy. The prepared dataset BANSpEemo in this research, where we used MFCC, Spectrogram, ZCR, and RMSE as feature extraction methods, which are common feature extraction techniques for speech recognition in machine learning. Following this, to show the accuracy or examine the performance, we applied various machine learning algorithms, namely SVM, LR, and MNB, with SVM performing the best, achieving a comparatively satisfactory accuracy level of 87.18%.

4.2.3. Confusion matrix analysis

A confusion matrix, commonly employed to evaluate a classification algorithm's efficacy, provides a summary of the model's predictions on a dataset and compares them with the actual labels. This matrix comprises four distinct entries:

- True positive (TP): events that are correctly classified as positive and that are genuinely positive.
- True negative (TN): events that are correctly classified as negative and that are genuinely negative.
- False positive (FP): when something is negative but is mistakenly classified as positive.
- False negative (FN): when something is positive but is mistakenly classified as negative.

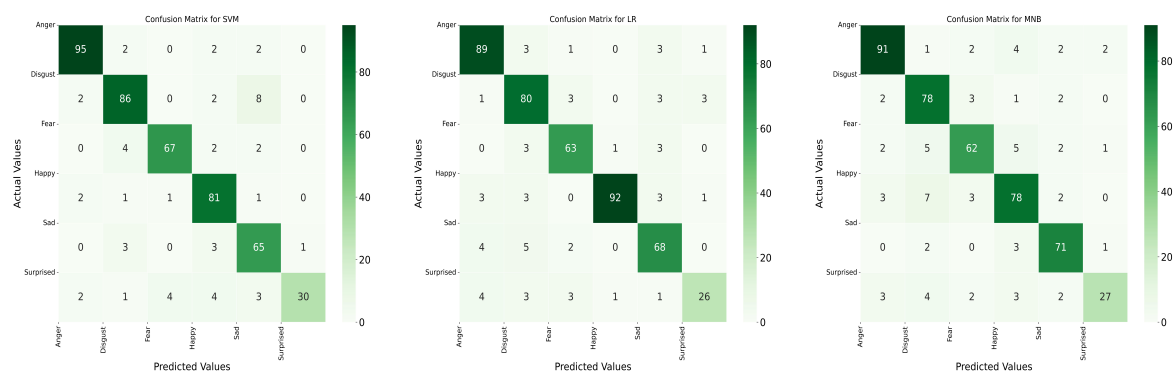


Figure 7. Confusion matrix for SVM, LR, and MNB (left to right)

The model's accuracy, precision, recall, and other performance indicators may all be assessed with the use of the confusion matrix. Figure 7 depicts the confusion matrices illustrating the performance of the chosen classification algorithms—SVM, LR, and multinomial Naive Bayes—in this research for evaluating the audio dataset.

4.2.4. ROC curve analysis

ROC curves, a simple yet effective visual representation of classifier accuracy, are used to evaluate the performance of these classifiers. ROC curves show the trade-off between the true positive rate and false-negative rate across various judgment thresholds. A common technique for turning a multi-class problem into a sequence of binary classification jobs is called OvR. In this strategy, every class is considered the positive class, and every other class is classified as the negative class. Following that, distinct ROC curves are produced, and AUC values are computed for every class. As a result, a collection of ROC curves and corresponding AUC values is obtained, with each class having its curve and score. A classifier is considered accurate when its curve reaches the upper-left corner, where the true positive rate is 1 (100%) and the false positive rate is 0. ROC curves for the multinomial Naive Bayes classifier, SVM, and logistic regression are presented in Figure 8. It is evident from the plots that the area under the curve (AUC) for SVM is greater than that of the other classifiers. Therefore, we can conclude that SVM is better suited for the classification of Bangla emotions.

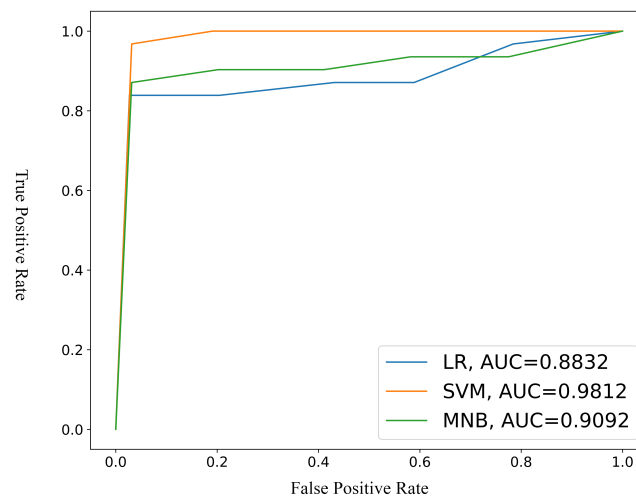


Figure 8. The ROC curve with AUC value for each model used in this research

5. CONCLUSION AND FUTURE DIRECTION

In this research, we introduced BANSpEmo, a new publically available dataset on emotion recognition from speech in the Bangla language. The collection contains 792 recordings of speech annotated for six different emotions. To assure annotation excellence, we employed several annotators with sufficient field expertise. To our knowledge, it's the fourth dataset especially created for recognizing emotions in Bangla speech. Furthermore, we presented classification performance outcomes on BANSpEmo for baseline machine learning algorithms, such as SVM, LR, and multinomial Naive Bayes. Our findings showed that SVM has the highest accuracy of 87.18% on the BanSpEmo dataset. These findings suggest that identifying emotions in Bangla voice data is difficult, particularly when using typical machine learning methods.

Although a professional soundproof room was not used during the creation of this audio dataset, the carefully selected sentences and the resulting dataset offer lexical diversity and linguistic benefits for future researchers. In addition to addressing a limit in the Bangla resource landscape, BANSpEmo establishes a standard for producing language-specific datasets in Bangla languages. We intend to stimulate additional research and innovations in this significant field of computing by making BANSpEmo available to the research community. We also plan to expand this dataset, which will benefit future researchers. After increasing the dataset size, we intend to apply deep learning algorithms to achieve higher accuracy and gain a deeper understanding of the underlying patterns.

ACKNOWLEDGEMENTS

This work was supported in part by the Center for Research, Innovation, and Transformation of Green University of Bangladesh.

REFERENCES

- [1] R. D. G. Ayon, Md. S. Rabbi, U. Habiba, and M. Hasana, "Bangla Speech Emotion Detection using Machine Learning Ensemble Methods," *Advances in Science, Technology and Engineering Systems Journal*, vol. 7, no. 6, pp. 70-76, 2022, doi: 10.25046/aj070608.
- [2] O. Sen *et al.*, "Bangla natural language processing: A comprehensive analysis of classical, machine learning, and deep learning-based methods," *IEEE Access*, vol. 10, pp. 38 999–39 044, 2022, doi: 10.1109/ACCESS.2022.3165563.
- [3] A. Sarkar, A. P. Sourav, and R. Ahmed, "Sentiment analysis in bengali text using nlp," Ph.D. dissertation, Brac University, 2023.
- [4] B. Sultana and K. A. Mamun, "Enhancing bangla-english code-mixed sentiment analysis with cross-lingual word replacement and data augmentation," in *2024 6th International Conference on Electrical Engineering and Information & Communication Technology (ICEEICT)*, 2024, pp. 652–657, doi: 10.1109/ICEEICT62016.2024.10534454.
- [5] M. Tareq, M. F. Islam, S. Deb, S. Rahman, and A. Al Mahmud, "Data-augmentation for bangla-english code-mixed sentiment analysis: Enhancing cross linguistic contextual understanding," *IEEE Access*, 2023, doi: 10.1109/ACCESS.2023.3277787.
- [6] T. Ahmed, S. F. Mukta, T. Al Mahmud, S. Al Hasan, and M. G. Hussain, "Bangla text emotion classification using lr, mnb and mlp with tf-idf & countvectorizer," in *2022 26th International Computer Science and Engineering Conference (ICSEC)*, 2022, pp. 275–280, doi: 10.1109/ICSEC56337.2022.10049341.

- [7] W. Fan, X. Xu, X. Xing, W. Chen, and D. Huang, "Lssed: a large-scale dataset and benchmark for speech emotion recognition," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 641–645, doi: 10.1109/ICASSP39728.2021.9414542.
- [8] S. R. Livingstone and F. A. Russo, "The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multi-modal set of facial and vocal expressions in north american english," *PloS one*, vol. 13, no. 5, p. e0196391, 2018, doi: 10.1371/journal.pone.0196391.
- [9] A. Asghar, S. Sohaib, S. Iftikhar, M. Shafi, and K. Fatima, "An urdu speech corpus for emotion recognition," *PeerJ Computer Science*, vol. 8, p. e954, 2022, doi: 10.7717/peerj-cs.954.
- [10] T. Wu, Y. Yang, Z. Wu, and D. Li, "Masc: A speech corpus in mandarin for emotion analysis and affective speaker recognition," in *2006 IEEE odyssey-the speaker and language recognition workshop*, 2006, pp. 1–5, doi: 10.1109/ODYSSEY.2006.248084.
- [11] J. Hintz, A. Wendemuth, and I. Siegert, "Cross-reliability benchmark test for preserving emotional content in speech–synthesis related datasets," in *Konferenz Elektronische Sprachsignalverarbeitung. TUDpress, Dresden*, 2023, pp. 64–71.
- [12] B. Nojavanasghari, T. Baltrušaitis, C. E. Hughes, and L.-P. Morency, "Emoreact: a multimodal approach and dataset for recognizing emotional responses in children," in *Proceedings of the 18th acm international conference on multimodal interaction*, 2016, pp. 137–144, doi: 10.1145/2993148.2993168.
- [13] A. Zadeh, Y. S. Cao, S. Hessner, P. P. Liang, S. Poria, and L.-P. Morency, "Cmu-moseas: A multimodal language dataset for spanish, portuguese, german and french," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, 2020, vol. 2020, p. 1801, doi: 10.18653/v1/2020.emnlp-main.141.
- [14] M. M. Billah, M. L. Sarker, and M. Akhand, "Kbes: A dataset for realistic bangla speech emotion recognition with intensity level," *Data in Brief*, vol. 51, p. 109741, 2023, doi: 10.1016/j.dib.2023.109741.
- [15] S. Sultana, M. Z. Iqbal, M. R. Selim, M. M. Rashid, and M. S. Rahman, "Bangla speech emotion recognition and cross-lingual study using deep cnn and blstm networks," *IEEE Access*, vol. 10, pp. 564–578, 2022, doi: 10.1109/ACCESS.2021.3136251.
- [16] R. K. Das, N. Islam, M. R. Ahmed, S. Islam, S. Shatabda, and A. M. Islam, "Banglaser: A speech emotion recognition dataset for the bangla language," *Data in Brief*, vol. 42, p. 108091, 2022, doi: 10.1016/j.dib.2022.108091.
- [17] P. Dhar and S. Guha, "A system to predict emotion from bengali speech," *International Journal of Mathematical Sciences and Computing (IJMSC)*, vol. 7, no. 1, pp. 26–35, 2021, doi: 10.5815/IJMSC.2021.01.04.
- [18] J. Devnath, S. Hossain, M. Rahman, H. Saha, M. A. Habib, and N. Sultan, "Emotion recognition from isolated bengali speech," *Journal of Theoretical and Applied Information Technology*, vol. 98, no. 10, pp. 1523–1533, 2020.
- [19] S. Sultana, M. S. Rahman, M. R. Selim, and M. Z. Iqbal, "Sust bangla emotional speech corpus (subesco): An audio-only emotional speech corpus for bangla," *Plos one*, vol. 16, no. 4, p. e0250173, 2021, doi: 10.1371/journal.pone.0250173.
- [20] M. M. Hassan, M. Raihan, M. M. Hassan, and A. K. Bairagi, "Bser: A learning framework for bangla speech emotion recognition," in *2024 6th International Conference on Electrical Engineering and Information & Communication Technology (ICEEICT)*, 2024, pp. 410–415, doi: 10.1109/ICEEICT62016.2024.10534493.
- [21] M. R. Islam, M. Akhand, and M. A. S. Kamal, "Bangla speech emotion recognition using 3d cnn bi-lstm model," in *International Conference on Machine Intelligence and Emerging Technologies*, 2022, pp. 539–550, doi: 10.1007/978-3-031-34619-4_42.
- [22] S. Sultana and M. S. Rahman, "Acoustic feature analysis and optimization for bangla speech emotion recognition," *Acoustical Science and Technology*, vol. 44, no. 3, pp. 157–166, 2023, doi: 10.1250/ast.44.157.
- [23] S. Aziz, N. H. Arif, S. Ahbab, S. Ahmed, T. Ahmed, and M. H. Kabir, "Improved speech emotion recognition in bengali language using deep learning," in *2023 26th International Conference on Computer and Information Technology (ICCIT)*, 2023, pp. 1–6, doi: 10.1109/ICCIT60459.2023.10441053.
- [24] R. Lotfian and C. Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings," *IEEE Transactions on Affective Computing*, vol. 10, no. 4, pp. 471–483, 2017, doi: 10.1109/TAFFC.2017.2736999.
- [25] J. Sager, R. Shankar, J. Reinhold, and A. Venkataraman, "Vesud: A crowd-annotated database to study emotion production and perception in spoken english," in *Interspeech*, 2019, pp. 316–320, doi: 10.21437/Interspeech.2019-1413.
- [26] S. Mandal, B. Das, and P. Mitra, "Shruti-ii: A vernacular speech recognition system in bengali and an application for visually impaired community," in *2010 IEEE students technology symposium (TechSym)*, 2010, pp. 229–233, doi: 10.1109/TECHSYM.2010.5469156.
- [27] M. G. Hussain, M. Rahman, B. Sultana, A. Khatun, and S. Al Hasan, "Classification of bangla alphabets phoneme based on audio features using mlpc & svm," in *2021 International Conference on Automation, Control and Mechatronics for Industry 4.0 (ACMI)*, 2021, pp. 1–5, doi: 10.1109/ACMI53878.2021.9528088.
- [28] M. G. Hussain, M. Rahman, B. Sultana, and Y. Shiren, "Banspemo: A bangla emotional speech recognition dataset," *arXiv preprint arXiv:2312.14020*, 2023, doi: 10.48550/arXiv.2312.14020.
- [29] M. M. Billah, L. Sarker, M. Akhand, and M. A. Samad Kamal, "Emotion recognition with intensity level from bangla speech using feature transformation and cascaded deep learning model," *International Journal of Advanced Computer Science & Applications*, vol. 15, no. 4, 2024, doi: 10.14569/ijacsa.2024.0150460.
- [30] A. C. Shruti, R. H. Rifat, M. Kamal, and M. G. R. Alam, "A comparative study on bengali speech sentiment analysis based on audio data," in *2023 IEEE International Conference on Big Data and Smart Computing (BigComp)*, 2023, pp. 219–226, doi: 10.1109/BigComp57234.2023.00043.

BIOGRAPHIES OF AUTHORS

Babe Sultana    was born in, Cox's Bazar, Bangladesh. She received her B.Sc. Degree in Computer Science and Engineering from Green University of Bangladesh (GUB) in 2018. At present, she is pursuing her MSc in CSE from United International University and working as a Lecturer, Dept. of CSE, Green University of Bangladesh. Also, her publication was about "Multi-mode Project Scheduling with Limited Resource and Budget Constraints" published in International Conference on Innovation in Engineering and Technology (ICIET) 27-28 December, 2018 and she got Best Paper Award and IEEE Best Paper Award on this conference. Her research interests include theory of optimization, natural language processing, machine learning, and renewable and sustainable energy. She can be contacted at email: babecse@gmail.com.



Md Gulzar Hussain    was born in Dinajpur City, Bangladesh. He just started his Ph.D. in the School of Software at Nanjing University of Information Science and Technology, Nanjing, China. He received his Master's degree (2024) from the School of Computer Science and Artificial Intelligence, Changzhou University, Changzhou, China, and his Bachelor of Science degree (2018) in Computer Science and Engineering from the Green University of Bangladesh, Dhaka, Bangladesh. He worked as a Lecturer in the Computer Science and Engineering Department, at Green University of Bangladesh, Dhaka, Bangladesh from 2019 to 2022. He is affiliated with IEEE as a student member and IAER as a professional member. His research interests include transfer learning, natural language processing, text mining, and topic modeling. He can be contacted at email: gulzar.ace@gmail.com.



Mahmuda Rahman    received her Master's in CS from Jahangirnagar University and B.Sc. in CSE from Green University of Bangladesh. She has received Vice Chancellor's Gold Medal award for her outstanding academic performance in 4th convocation of Green University of Bangladesh. She worked as a Lecturer, Dept. of CSE, Green University of Bangladesh from 2019 to 2023. At present, she has been working as a Lecturer (ICT), Mohammadpur Preparatory School and College, Dhaka, Bangladesh. Her research interest includes natural language processing, medical image processing, and machine learning. She can be contacted at email: mahmuda.mpsc@gmail.com.